An Investigation into the Validity of Some Metrics for Automatically Evaluating Natural Language Generation Systems

Ehud Reiter*
University of Aberdeen

Anja Belz**
University of Brighton

There is growing interest in using automatically computed corpus-based evaluation metrics to evaluate Natural Language Generation (NLG) systems, because these are often considerably cheaper than the human-based evaluations which have traditionally been used in NLG. We review previous work on NLG evaluation and on validation of automatic metrics in NLP, and then present the results of two studies of how well some metrics which are popular in other areas of NLP (notably BLEU and ROUGE) correlate with human judgments in the domain of computer-generated weather forecasts. Our results suggest that, at least in this domain, metrics may provide a useful measure of language quality, although the evidence for this is not as strong as we would ideally like to see; however, they do not provide a useful measure of content quality. We also discuss a number of caveats which must be kept in mind when interpreting this and other validation studies.

1. Introduction

Evaluation is becoming an increasingly important topic in Natural Language Generation (NLG), as in other fields of computational linguistics. Many NLG researchers are impressed by the BLEU evaluation metric (Papineni et al. 2002) in Machine Translation (MT), which has allowed MT researchers to quickly and cheaply evaluate the impact of new ideas, algorithms, and data sets. BLEU and related metrics work by comparing the output of an MT system to a set of reference translations (human translations of the source text), and in principle this kind of evaluation could be done with NLG systems as well. As in other areas of NLP, the advantages of automatic corpus-based evaluation are that it is potentially much cheaper and quicker than human-based evaluation, and that it is repeatable. Indeed, NLG researchers have used BLEU in their evaluations for some time (Langkilde 2002; Habash 2004).

The use of such automatic evaluation metrics is, however, only sensible if they are known to be correlated with the results of reliable human-based evaluations. Although a number of previous studies have analyzed correlations between human judgments

* Department of Computing Science, University of Aberdeen, UK. E-mail: e.reiter@abdn.ac.uk.
** Natural Language Technology Group, University of Brighton, UK. E-mail: A.S.Belz@brighton.ac.uk.

Submission received: 23 March 2007; revised submission received: 6 October 2008; accepted for publication: 29 December 2008.

and automatic evaluation metrics in machine translation and document summarization (Doddington 2002; Papineni et al. 2002; Lin and Hovy 2003), much less is known about how well automatic metrics correlate with human judgments in NLG. In this article we present two empirical studies of how well BLEU and various other corpus-based metrics agree with human judgments, when evaluating the outputs of several NLG systems that generate texts which describe changes in the wind (for weather forecasts). We also discuss several caveats that need to be kept in mind when interpreting our study and perhaps other validation studies of automatic metrics as well.

2. Background: Evaluation in NLG and Related Fields

As Hirschman (1998), Mellish and Dale (1998), and others have pointed out, evaluations can be used for many purposes, and different evaluations are often needed for different stakeholders. For example, the BabyTalk project at Aberdeen (Portet et al. 2009), which is attempting to create a set of NLG systems which can generate textual summaries of clinical data about babies in a neonatal intensive care unit (NICU), is a collaboration between medical researchers, psychologists, computer scientists, and a commercial software house. Each of these groups has its own evaluation agenda:

- The *medical researchers* want to know if BabyTalk is medically effective. To evaluate this, they ideally would like to do a study similar to Cunningham et al.'s (1998) evaluation of the effectiveness of a visualization system in an intensive care unit; that is, deploy BabyTalk in a hospital, use it for half of the children in a ward, and determine if there is any difference in outcome (e.g., mortality) between the children in the BabyTalk group and the control group.
- The *psychologists* want to understand the effectiveness of textual presentation of information for decision support. To evaluate this, they would like to do a study similar to Law et al. (2005); that is, show medical subjects textual summaries (as well as standard graphical visualizations as a control) in a controlled "off-ward" context, ask them to make a treatment decision, and compare this decision against a gold standard.
- The *computer scientists* want to know if BabyTalk is effective (under either of these measures). They also would like to conduct evaluations throughout the project, so as to assess whether their development efforts are making the system better or worse; the stakeholders, in contrast, would be satisfied with a single evaluation at the end of the project.
- The *software house* would like to know if BabyTalk would be commercially profitable. This partially depends on medical effectiveness (see previous point), which determines the demand for the system. But it also depends on how expensive it is to develop and support BabyTalk; from this perspective the company is especially interested in evaluations of the cost of adapting/porting BabyTalk to different hospitals in the NICU domain in the short term, and to different medical domains in the longer term. See Harris (2008) for a commercial perspective on medical NLG systems.

All of these stakeholders are interested in evaluations which assess the quality and effectiveness of generated texts; such evaluations are the focus of our article. The software house and the computer scientists are also interested in engineering-cost evaluations; although this is a very important topic, we will not discuss it here: a separate article would be needed to do justice to this topic.

2.1 Evaluation in NLG

The quality of texts generated by NLG systems has been evaluated in many different ways in the past, most of which can be classified as evaluations based on task performance, human judgments and ratings, or comparison to corpus texts using automatic metrics.

2.1.1 Task-Based Evaluation. Task-based evaluations involve directly measuring the impact of generated texts on end users; these are extrinsic evaluations (Spärck Jones and Galliers 1995), and typically involve techniques from psychology or from an application domain such as medicine. One of the first task-based evaluations of an NLG system was done by Young (1999), who generated instructional texts using four different algorithms, asked subjects to carry out the instructions, and then measured how many mistakes they made. Although task performance is the most common measure used in task-based evaluations in NLG, other measures can also be used. For example, Carenini and Moore (2006) evaluated the impact of persuasive texts (in a house-selling context) by seeing how users ranked houses in a hot list; and Di Eugenio, Glass, and Trolio (2002) evaluated the impact of adding an NLG component to an intelligent tutoring system by measuring learning gain.

We have been involved in a number of task-based evaluations of NLG systems and components. STOP (Reiter, Robertson, and Osman 2003), which generates personalized smoking-cessation letters, was evaluated on the basis of medical effectiveness; we sent a group of 2,000 smokers either STOP-generated letters or one of two kinds of control letters, and measured how many smokers in each group managed to quit smoking. BT45 (Portet et al. 2009) (which is one of the BabyTalk systems) was evaluated for its decision-support effectiveness, using the "psychologist" methodology described earlier (van der Meulen et al. 2009). SKILLSUM (Williams and Reiter 2008), which generates feedback reports from literacy assessments, was evaluated on the basis of educational effectiveness; we gave 200 assessment takers either SKILLSUM texts or control texts, and measured whether they increased the accuracy of self-assessments of their literacy skills. We also evaluated several referring-expression generation algorithms by conducting experiments in which participants were presented with generated referring expressions and asked to identify the target referent (Belz and Gatt 2007; Gatt, Belz, and Kow 2008, 2009); these were carried out in conjunction with shared-task events organized under the Generation Challenges initiative (Generation Challenges is further discussed in Section 2.1.4).

Task-based evaluations have traditionally been regarded as the most meaningful kind of evaluation in NLG, especially in contexts where the evaluation needs to convince people in other communities (such as psychologists and doctors). However, they can be expensive and time-consuming. The STOP evaluation cost UK£75,000, and required 20 months to design, carry out, and analyze; the SKILLSUM and BT45 evaluations (which are perhaps more typical) cost about UK£20,000 over six months. The referring-expression identification experiments were cheaper (less than UK£1,000 each, not counting data and system creation), because they involved smaller numbers of subjects, and evaluated system components in laboratory-based settings, rather than by means of systems deployed in the real world.

In addition to monetary and time costs, all of these evaluations also depended on goodwill from participants, in most cases busy domain experts who used their own standing in their community to arrange access to subjects and otherwise facilitate the evaluation. Such goodwill in itself is a scarce resource which must be used with care.

2.1.2 Evaluations Based on Human Ratings and Judgments. Another way of evaluating an NLG system is to ask human subjects to rate generated texts on an *n*-point rating scale; this is an intrinsic form of evaluation (Spärck Jones and Galliers 1995). This methodology was first used in NLG by Lester and Porter (1997), who asked eight domain experts to each rate 15 texts on a number of different dimensions: overall quality and coherence, content, organization, writing style, and correctness. Some of the texts were human-written and some were computer-generated, but the judges did not know the origin of specific texts they read. Many more such evaluations have been performed since, often with fewer dimensions. For example, Binsted, Pain, and Ritchie (1997) evaluated a joke-generation system by asking children to rate the funniness of texts on a 5-point scale; and Walker, Rambow, and Rogati (2002) evaluated the SPoT sentence-planning system by asking human subjects to rate the overall quality of generated texts on a 5-point scale. A variation of this technique is to show subjects different versions of a text, and ask them which one they prefer. For example, the SumTime weather-forecast generator was evaluated by showing subjects both human corpus texts and computer-generated texts, and asking which they preferred (Reiter et al. 2005).

Evaluations based on human ratings and judgments are currently probably the most popular way of evaluating NLG systems, perhaps in part because such evaluations tend to be significantly quicker and cheaper to carry out than task-based evaluations, and do not require as much support from domain experts. For example, the previously mentioned evaluation of SumTime was carried out in two months without any external research grant funding.

In addition to resource issues, another reason why some researchers prefer evaluations based on human ratings over task-based evaluations is that task-based evaluations need to focus on a very specific task, and performance on this task may not correlate with performance on other tasks. For example, as mentioned previously, the medical researchers in BabyTalk would like to conduct a medical effectiveness evaluation, which involves operationally deploying systems in a real ward and measuring impact on patient outcome. However, for ethical reasons such an experiment cannot be carried out until we have good evidence that the BabyTalk systems are effective, which is not yet the case. We carried out an off-ward task-based evaluation of BT45 using the "psychologist" methodology (van der Meulen et al. 2009), and we would like to think that the results of this evaluation would correlate with the results of a medical effectiveness evaluation. However, we do not have any empirical evidence that this is the case, and certainly there are major differences between the off-ward and on-ward contexts (for example, doctors in the off-ward experiment could not visually observe the babies, which is a very important information source when doctors are actually caring for a baby in a hospital ward). From this perspective, an argument can be made that asking doctors to explicitly rate the medical usefulness of the texts might tell us as much about their genuine medical effectiveness as our off-ward task-based evaluation.

Last but not least, it is not always possible to conduct meaningful task-based evaluations of some NLG systems. For example, it is unclear how to evaluate the overall quality of jokes produced by a humor generation system other than by asking for human ratings (Binsted, Pain, and Ritchie 1997), although one can perform a task-based evaluation of the educational impact of humor generation software (Black et al. 2007), or (more speculatively) perhaps evaluate the psychological impact of a joke by monitoring facial expressions and laughter (which is a non-task-based extrinsic evaluation).

Little is known about how well human ratings of texts produced by NLG systems correlate with task-effectiveness measures. Law et al. (2005), who worked in the same domain (NICU) as BabyTalk, conducted an off-ward decision-support evaluation which compared human-written text summaries and graphical visualizations of clinical data. They found that subjects preferred the visualizations, but were more likely to make correct decisions from the text summaries. It is unclear whether this is because subjects had inappropriate preferences, or because there was a big difference between genuine medical effectiveness and off-ward decision-support effectiveness (as mentioned earlier). The only studies we are aware of which examined how well human judgments predict task-effectiveness of computer-generated texts occurred in the recent Generation Challenges evaluations of referring expression generation, which measured the correlations between human assessments of language quality and adequacy of content with task-performance measures (referent identification time and accuracy) (Gatt, Belz, and Kow 2009). The results revealed a strong and highly significant correlation between human judgments of content adequacy and identification accuracy; there was also a significant inverse correlation between human judgments of language quality and identification speed (i.e., those systems that tended to be judged more fluent by the human assessors also tended to have shorter identification times).

2.1.3 Evaluations Based on Automatic Metrics which Compare Computer-Generated Texts to Human-Authored Corpus Texts. In recent years there has been growing interest in evaluating NLG texts by comparing them to a corpus of human-written reference texts, using automatic metrics such as string-edit distance, tree similarity, or BLEU (Papineni et al. 2002); this is another type of intrinsic evaluation. Such evaluations have been used by Bangalore, Rambow, and Whittaker (2000) and Marciniak and Strube (2004), for example. Langkilde (2002) evaluated an NLG system by parsing texts from a corpus, feeding the parser output to her NLG system, and then comparing the generated texts to the original corpus texts. Similar "corpus regeneration" evaluations have since been used by a number of other researchers (Callaway 2003; Zhong and Stent 2005; Cahill and van Genabith 2006). Corpus-based evaluation has been especially popular in the evaluation of surface realizers. This may be because the most important attribute of many realizers is grammatical coverage and robust handling of special and unusual cases, and corpus-based techniques are well suited to evaluating this. Also, the range of acceptable outputs can be smaller in realizer evaluations because content, microplanning, and (in some cases) lexical choices do not vary; this means there is less concern about reference texts not adequately covering the solution space.

Automatic corpus-based evaluations are appealing in NLG, as in other areas of NLP, because they are relatively cheap and quick to do if a corpus is available, do not require support from domain experts, and are repeatable. However, their use in NLG is controversial, at least when evaluating systems as a whole instead of just surface realizers, because many people are concerned that the results of such evaluations may not be meaningful. For example Reiter and Sripada (2002) point out that corpus texts are often not of high enough quality to form good reference texts; and Scott and Moore (2007) express concern that metrics will not be able to evaluate many important linguistic properties such as information structure.

A more general concern is that automatic metrics based on comparison to reference texts measure how well a text matches what *writers* do, whereas most human evaluations (task or judgment-based) measure the impact of a text on *readers*. Because writers do not always produce optimal texts from a reader's perspective (Oberlander 1998; Reiter et al. 2005), a metric which is a good evaluator of how likely it is that a text has been written by a human writer is not necessarily a good predictor of how effective and useful the text is from the perspective of a human reader. Of course automatic metrics do not need to be writer-based. Indeed, some reader-based automatic metrics, such as the Flesch score (Flesch 1949) (based on average sentence and word length), are widely used as practical tools to help writers, but such metrics have not been widely used to evaluate NLP systems.

An important practical consideration is that corpus-based evaluations require a corpus of human-written reference texts; BLEU-like metrics in fact work best when the reference text corpus contains several reference texts for the same input, written by different authors. If reference texts have to be created specifically for an evaluation, this can be an expensive endeavor. In the BabyTalk domain, for example, it can take an experienced clinician several hours to write a corpus text from the raw data; hence creating a corpus of 100 reference texts in this domain could require 2–3 months effort by a clinician (as they do not create such reports in the course of their normal work). Getting this much time from an expert doctor or nurse would be difficult unless a very strong case could be made for the utility of the evaluation.

2.1.4 Other Validation Studies. In recent years some validation studies which examine correlations between automatic metrics and human evaluations in NLG have been carried out. The first such study we are aware of is Bangalore, Rambow, and Whittaker (2000), who looked at string-edit and tree-edit metrics (this work predates BLEU and ROUGE) using a small number of manually simulated system "outputs." Probably the most similar study to our work is that by Stent, Marge, and Singhai (2005), who examined the correlation between human judgments and several automatic metrics when evaluating computer-generated paraphrases; this is further discussed in Section 3.3.3.

Very recently some validation studies have been done in the context of the Generation Challenges initiative for shared tasks in NLG, by evaluating systems entered in the shared task using automatic metrics, human ratings, and task-based evaluation, and analyzing correlations between these. For example Belz and Gatt (2008) analyzed correlations between several automatic evaluation metrics and task performance in a referring-expression generation task; they found that there was no significant correlation between any of the automatic metrics they looked at (which included specialized metrics for the reference task as well as BLEU and ROUGE) and their task-based measures of effectiveness, such as how long it took human subjects to identify objects from a referring expression, and how many mistakes the subjects made. However the different automatic metrics they examined did tend to correlate with each other, as did the different measures of task performance. In general shared tasks offer a promising context for validation studies, and we hope that future Generation Challenges events will continue to provide data on how well automatic metrics correlate with human-based evaluations.

2.2 Insights from Evaluations in Other Areas of NLP

Of course, evaluation and experimentation are crucial to all fields of NLP; here we look at insights from two other NLP subfields which need to evaluate the quality of texts: machine translation and document summarization.

2.2.1 Evaluation in Machine Translation. There is a rich literature in MT evaluation, including a number of specialist workshops on this topic; as in NLG, there is also considerable interest in using shared-task events to provide data about how well different evaluation techniques correlate with each other (Callison-Burch et al. 2008). From an NLG perspective, the most surprising aspect of current MT evaluation is the dominance of BLEU and other automatic corpus-based metrics (Callison-Burch, Osborne, and Koehn 2006). BLEU was first proposed as a supplement (the U in BLEU stands for "understudy") for human evaluation (Papineni et al. 2002), but it is now routinely used as the main technique for evaluating research contributions. It is accepted and indeed the norm for an article on MT in *Computational Linguistics* to report evaluations that are solely based on automatic corpus-based metrics; this is not the case in NLG, where human evaluations are expected at least in high-prestige venues.

We are not aware of any studies in MT that have tried to correlate BLEU-like metrics with the results of task-effectiveness studies. Although a number of studies have analyzed the correlation between BLEU-type metrics and human judgments, most of these have used human judgments from NIST MT evaluations. Human judgments in most of these evaluations were solicited from monolingual subjects who were asked to compare the output of MT systems to a single reference translation, without any context; also in many of these studies the subjects were asked to assess individual sentences or even phrases, not complete texts (Doddington 2002). As Coughlin (2003) and others have pointed out, it is not clear that human judgments solicited in this way would match the judgments of bilingual subjects who were shown complete source and MT texts, and asked to evaluate the quality of the translation in a specific real-world context. Papineni et al. (2002) in fact found that BLEU scores were more highly correlated with human judgments from monolingual subjects than human judgments from bilingual subjects.

In any case, regardless of the effectiveness of BLEU as an MT evaluation metric, another issue is whether an MT evaluation technique can in general be expected to work as an NLG evaluation technique. There are some obvious differences between MT systems and NLG systems; for example:

- *Content determination:* NLG systems need to decide on what information should be communicated in a text, as well as how this information is linguistically expressed; MT systems generally do not have to perform content determination.
- *Linguistic variety:* Many NLG systems produce text that is fairly simple from a linguistic perspective (partially because many NLG users prefer such texts); MT systems, in contrast, usually need to produce linguistically complex texts.
- *Genre/domain:* Most applied NLG systems (with some exceptions) try to generate high-quality texts in a limited domain and genre such as marine weather forecasts; MT systems, in contrast, typically generate lower-quality texts in a broad text category such as newspaper articles.

These differences presumably need to be considered when deciding whether it makes sense to use an MT evaluation technique in NLG. For example, there is no reason to expect MT evaluation techniques to be useful for evaluating NLG content determination, since MT systems do not perform this task. Also, MT evaluation techniques which work well when evaluating less-than-human-quality texts from an MT system may not necessarily work well when evaluating human-quality texts produced by an NLG system.

2.2.2 Evaluation in Document Summarization. Another branch of NLP which requires the evaluation of textual documents is document summarization. From an evaluation perspective, an important difference between MT and summarization is that summarization evaluations have placed much more emphasis on content determination. Perhaps in part because of this, the summarization community places more emphasis on human evaluations. Although there are automatic corpus-based metrics for summarization such as ROUGE (Lin and Hovy 2003), they do not seem to dominate summarization evaluation in the same way that BLEU-type metrics dominate MT evaluation.

The main summarization evaluation technique in the NIST TAC 2008 summarization track is the pyramid technique (Nenkova and Passonneau 2004), which is a structured human-based evaluation, based on asking human judges to identify 'summarization content units' (SCU) in model and system-generated summaries, and measuring how many SCUs from the model summaries occur in the system summary. This is an interesting technique for evaluating content, and might be worth investigating for evaluating content determination in NLG systems.

In terms of validation, a number of studies have claimed that ROUGE correlates with human ratings, for example Lin and Hovy (2003) and Dang (2006). Dorr et al. (2005) checked if ROUGE scores correlated with task effectiveness; they did not find a strong correlation.

2.3 Summary

In summary, evaluation of NLG texts in the past has primarily been done using human subjects, either by measuring the impact of texts on task performance, or by asking subjects to rate texts. However, a growing number of NLG researchers are using automatic metrics to evaluate their systems, perhaps inspired by the popularity of automatic metrics in other areas of NLP which involve evaluating output texts, most notably machine translation and document summarization. This use of metrics assumes that they correlate with human-based evaluations. A number of studies in machine translation and document summarization have shown that some automatic metrics correlate with human ratings; however we are not aware of any studies in these areas which have shown any metric to strongly correlate with task performance. Fewer validation studies have been carried out in NLG, although this is beginning to change as researchers place more importance on such studies.

3. Our Experiments

Given the growing interest in using automatic evaluation metrics such as BLEU in NLG, we decided to carry out some experiments to determine how well such metrics predicted the results of human judgments. As in other such studies, we did this by evaluating a number of systems with the same input/output functionality, using different evaluation techniques, and then analyzing the correlation between the techniques.

One potential weakness of our experiments was that we did not look at correlations with task-effectiveness evaluations. This was because we did not have the resources (money and domain-expert goodwill) to conduct a task-based evaluation. This issue is further discussed in Section 4.2.

3.1 Domain and Systems

Our work was done in the domain of computer-generated weather forecasts. This is one of the most popular applications of NLG (Goldberg, Driedger, and Kittredge 1994; Coch 1998; Reiter et al. 2005), and several NLG weather-forecast systems have been fielded and used. Weather forecast generation is probably the closest that NLG comes to a "standard" application domain, and hence seems a good choice for validation studies from this perspective.

On the other hand, though, one could also argue that weather-forecast generators are atypical in that the language they generate tends to be very simple, even by the standards of NLG systems: very limited syntax (which differs from conventional English), very small vocabulary, no real text structure above the sentence level, and so on. Hence it is not clear to what degree results obtained in this domain will generalize to other domains with less simple language and content (such as BabyTalk); this is further discussed in Section 4.1.

From a practical perspective, the great advantage of the weather forecast domain was that we had access to a number of systems, built using different NLG technologies, with the same input/output functionality; at the time (2006) there was no other domain where this was the case. This situation is changing, because of the emergence of shared-task events in NLG such as Generation Challenges (see Section 2.1.4).

3.1.1 SumTime Systems and SumTime-Meteo Corpus. In particular, we based our experiments on the SumTime system (Sripada et al. 2004; Reiter et al. 2005) and its associated SumTime-Meteo corpus (Sripada et al. 2003), which were developed at Aberdeen. SumTime generates textual weather forecasts from numerical forecast data for offshore oil rigs. It has two modules: a content-determination module that determines the content of the weather forecast by analyzing the numerical data using linear segmentation and other data analysis techniques; and a microplanning and realization module which generates texts based on this content by choosing appropriate words, deciding on aggregation, enforcing the sublanguage grammar, and so forth. SumTime generates very high-quality texts; in some cases forecast users believe SumTime texts are better than human-written texts (Reiter et al. 2005; see also Table 4 of this paper). The SumTime system has been used operationally to produce draft weather forecasts; these are post-edited by meteorologists before they are released to end users (Sripada et al. 2004).

SumTime is a knowledge-based NLG system. Although its design was informed by corpus analysis (Reiter, Sripada, and Robertson 2003), the system is composed of manually authored rules and code.

The SumTime project also created a corpus and data set, called SumTime-Meteo (Sripada et al. 2003). This consists of a corpus of 1,045 weather forecasts written by professional forecasters, and the numerical predictions of wind, temperature, and so forth, that forecasters examined when they wrote the forecasts. For wind descriptions only, the corpus also contains simple content representations containing information about wind speed and direction, time of day, and position in forecast (we call these "content tuples"). The content tuples were created by parsing the corpus texts and extracting the relevant information (Reiter and Sripada 2003), and are similar to the representations produced by the SumTime content-determination system. Figures 1, 2, and 3 show an extract from a numerical data file, an extract from the corresponding human-written forecast, and the content tuples derived from the human text.

day/hour	wind direction	avg wind speed	max (gust) wind speed
05/06	SSW	18	22
05/09	S	16	20
05/12	S	14	17
05/15	S	14	17
05/18	SSE	12	15
05/21	SSE	10	12
06/00	VAR	6	7

Figure 1
Extract from meteorological data file for 05-10-2000 (morning forecast).

```
FORECAST 06-24 GMT, THURSDAY,         05-Oct   2000

WIND(KTS)      CONFIDENCE: HIGH
  10M:         SSW 16-20 GRADUALLY BACKING SSE THEN FALLING
               VARIABLE 04-08 BY LATE EVENING
...
```

Figure 2
Extract from corpus (human) forecast for 05-10-2000 (morning forecast).

index	wind direction	min wind speed	max wind speed	time
1	SSW	16	20	0600
2	SSE	-	-	-
3	VAR	04	08	2400

Figure 3
Content tuples extracted from the forecast in Figure 2.

In addition to the main SumTime system, two other generation methods were developed in the SumTime project and used in the experiments described here. SumTime-Hybrid uses the SumTime microplanner/realizer to generate text from the corpus-derived content tuples (Figure 3). In other words, it combines human content-determination with SumTime microplanning and realization. The other method is an algorithm which is based on a spreadsheet and flowchart which one of the forecasters gave to the SumTime team at the beginning of the project (Reiter, Sripada, and Robertson 2003, page 499); a simplified version of this algorithm is presented in Figure 4. We did not implement this algorithm as software, but we manually executed it for a number of forecasts. We refer to it below as the **Template system** since the linguistic part of the flowchart was based on template-filling.

3.1.2 pCRU Generators for the SumTime-Meteo Domain. Independently of the SumTime Project, we created a range of statistical generators for the SumTime-Meteo domain using *p*CRU generation (probabilistic context-free representational underspecification) (Belz 2008). These took content tuples as input (as in Figure 3), not meteorological data files (as in Figure 1); in other words, they did not perform content determination.

*p*CRU is a probabilistic language generation framework that was developed with the aim of providing the formal underpinnings for creating narrow-domain, applied NLG systems that are driven by comprehensive probabilistic models of the entire generation

```
Start text with direction, speed (5 kt range around actual value) from the first entry
  in the data file
For each subsequent data file entry
        If direction has changed 45 degrees or more since last mentioned direction, or
          direction has changed at all, and speed is greater than 15 kts, then
          add the following phrase to the text
            • veering if direction change is clockwise, backing otherwise
            • new direction
            • new speed (5 kt range around value) if speed has changed by 5 kts or more
            • time phrase (from fixed table which maps numeric time to a phrase)
        Else if speed has changed by 5 kts or more since last mentioned speed, then
          add the following phrase to the text
            • becoming
            • new speed (5 kt range around actual value)
            • time phrase (from fixed table which maps numeric time to a phrase)
        end if
end for
```

Figure 4
Template algorithm (simplified by removing special cases).

space. NLG systems are modeled as sets of generation rules that apply transformations to representations. The basic idea in *p*CRU is that as long as the generation rules are all of the form $relation(arg_1, ...arg_n) \rightarrow relation_1(arg_1, ...arg_p)$ ... $relation_m(arg_1, ...arg_q)$, $m \geq 1, n, p, q \geq 0$, then the set of all generation rules can be seen as defining a context-free language and a single probabilistic model can be estimated from raw or annotated text to guide the generation processes.

*p*CRU uses straightforward context-free technology in combination with underspecification techniques, to encode a **base generator** as a set of expansion rules G. The *p*CRU **decision-maker** is created by estimating a probability distribution over the base generator, as follows:

1. *Convert corpus into multi-treebank:* Determine for each sentence all (left-most) derivation trees licensed by the base generator's rules, using maximal partial derivations if there is no complete derivation tree; annotate the (sub)strings in the sentence with the derivation trees, resulting in a set of *generation trees* for the sentence.
2. *Train base generator:* Obtain frequency counts for each individual generation rule from the multi-treebank, adding $1/n$ to the count for every rule, where n is the number of alternative derivation trees; convert counts into probability distributions over alternative rules, using add-1 smoothing and standard maximum likelihood estimation.

The resulting probability distribution is used in one of the following three ways to control generation.

1. *Viterbi generation:* Do a Viterbi search of the generation forest for a given input, which maximizes the joint likelihood of all decisions taken in the generation process.

2. *Greedy generation:* Make the single most likely decision at each choice point (rule expansion) in a generation process.

3. *Greedy roulette-wheel generation:* Base decisions on a non-uniform random distribution proportional to the likelihoods of alternatives.

We also implemented two baseline *p*CRU systems, both of which ignore *p*CRU probabilities: the **random** mode, which randomly selects generation rules; and the **n-gram** mode, which generates the set of alternatives and selects the most likely one according to an *n*-gram language model (Langkilde and Knight 1998).

Combining the *p*CRU, SUMTIME, and Template systems gave us a set of systems which had the same target functionality, but attempted to achieve it using quite different NLG techniques and technologies. Tables 1 and 2 show examples of texts produced by humans and our systems. The human texts include reference texts for automatic metrics (see Section 3.3) as well as the corpus texts.

Note that we could not use other marine weather-forecast generators in our experiments, such as FOG (Goldberg, Driedger, and Kittredge 1994), because they use different inputs (that is, different numerical weather prediction models), and they produce outputs targeted at different audiences (e.g., FOG forecasts are intended for mariners in general, whereas SUMTIME forecasts are targeted at the offshore oil industry).

3.2 Human Evaluations

We conducted two experiments where we asked human subjects to rate texts produced by our different marine weather-forecast generators. The main difference was that the first experiment focused on evaluating linguistic quality, and only looked at texts with the same information content. The second experiment also evaluated content quality, and used texts that varied in content as well as in linguistic expression. We also changed the experimental design in the second experiment, based on our experiences in the first experiment.

3.2.1 First Human Evaluation. In Experiment 1 (the main results of which we reported previously in Belz and Reiter [2006]), we focused on the content-to-realization mapping, so we restricted ourselves to systems which generated texts from content tuples (Figure 3) (SUMTIME-Hybrid and the *p*CRU systems). We also included the corresponding texts from the SUMTIME-METEO corpus.

We used a randomly selected subset of 21 forecast dates from the SUMTIME-METEO corpus. We restricted ourselves to morning forecasts (half the corpus), as these are based on a single data file (evening forecasts are based on two data files), and to the first wind description in a forecast, as subsequent wind descriptions have the added constraint of being consistent in form and content with earlier wind descriptions. For each of these dates, we obtained seven texts: the corpus text, the texts produced by the previously mentioned systems, and one of the reference texts used by the automatic metrics (Section 3.3.1).[1] This gave us a total of 147 texts.

[1] We wanted to obtain ratings for reference texts to check that these were regarded as reasonable by the human subjects. This was indeed the case, but we do not report the ratings of the reference texts here, because we do not have permission to do so. Also we cannot use human ratings of reference texts in the correlation studies reported in Section 3.3, because automatic metrics cannot be used to evaluate their own reference texts.

Table 1
Texts produced for 5 Oct 2000, from content tuples in Figure 3.

Human texts:	
Corpus	SSW 16-20 GRADUALLY BACKING SSE THEN FALLING VARIABLE 4-8 BY LATE EVENING
Reference 1	SSW'LY 16-20 GRADUALLY BACKING SSE'LY THEN DECREASING VARIABLE 4-8 BY LATE EVENING
Reference 2	SSW 16-20 GRADUALLY BACKING SSE BY 1800 THEN FALLING VARIABLE 4-8 BY LATE EVENING
Reference 3	SSW 16-20 GRADUALLY BACKING SSE THEN FALLING VARIABLE 04-08 BY LATE EVENING
System-generated texts:	
ST-Hybrid	SSW 16-20 GRADUALLY BACKING SSE THEN BECOMING VARIABLE 10 OR LESS BY MIDNIGHT
*p*CRU-greedy	SSW 16-20 BACKING SSE FOR A TIME THEN FALLING VARIABLE 4-8 BY LATE EVENING
*p*CRU-roulette	SSW 16-20 GRADUALLY BACKING SSE AND VARIABLE 4-8
*p*CRU-2gram	SSW 16-20 BACKING SSE VARIABLE 4-8 LATER
*p*CRU-random	SSW 16-20 AT FIRST FROM MIDDAY BECOMING SSE DURING THE AFTERNOON THEN VARIABLE 4-8

Table 2
Texts produced for 6 Oct 2000, directly from numerical wind data.

Human texts:	
Corpus	N-NW 12-18 BACKING / EASING SSW LESS THAN 08 BY END OF PERIOD
Reference 1	N-NW 12-17 DECREASING 10 OR LESS BY LATE AFTERNOON AND BACKING SW LATER
Reference 2	NW-NNW 12-15 KNOTS DECREASING AND BACKING TO BE WSW 05-08 BY EVENING AND VARIABLE LESS THAN 05 KNOTS BY MIDNIGHT
Reference 3	W-NNW 12-16 BACKING AND EASING THROUGH LATE AFTERNOON / EARLY EVENING SW-SSW LESS THAN 10
System-generated texts:	
SUMTIME	NNW 12-17 EASING 10 OR LESS BY MID AFTERNOON THEN BACKING WSW BY LATE AFTERNOON AND SSW BY MIDNIGHT
Template	NNW 13-18 BECOMING 08-13 IN THE MID-AFTERNOON BACKING WSW IN THE EARLY EVENING BECOMING 03-08 DURING THE NIGHT BACKING SSW 00-05 AROUND MIDNIGHT

For our human evaluators, we recruited nine people with experience reading forecasts for offshore oil rigs ('experts'). Note that these were experienced forecast readers, not forecast writers. We also recruited 21 people with no experience in reading forecasts for offshore oil rigs ('non-experts'). The reason for including non-experts was that we wanted to see if ratings by non-experts were similar to ratings by experts (as non-experts are often much easier to recruit for experiments). None of the subjects had a background in NLP, and all were native speakers of English.

Subjects were shown forecast texts from all the content-to-text generators, and from the corpus, and asked to give each text a single score on a scale of 0 to 5, which was explained as reflecting *readability, clarity, and general appropriateness*. Experts (only) were also shown the numerical weather data that the forecast text was based on. Subjects were *not* shown reference texts, as is often done in MT evaluations (Section 2.2.1). The experiment was done over the World Wide Web, at a time and place convenient to the subjects.

All subjects were shown two practice examples at the beginning of the test which were not included in the analysis. Expert subjects were then shown one randomly selected text for 18 of the dates. The non-experts were shown 21 forecast texts, in a Repeated Latin Squares experimental design, where each of the 147 texts was rated by three subjects.

The average scores assigned by experts and non-experts are shown in Table 3. There was good correlation between experts and non-experts: Pearson's $r = 0.874$ (p = 0.011, one-tailed). Experts and non-experts also agreed about relative rankings, except that experts rank *p*CRU-greedy second and the corpus texts third, whereas the non-experts have these the other way around.

To determine if any of the differences were statistically significant, we used SPSS's General Linear Model (GLM), with rating as the dependent variable, and generator, subject, and forecast date as independent variables; we used a post hoc Tukey HSD test to identify significant differences between individual systems. This analysis is essentially equivalent to normalizing (adjusting) the ratings to remove differences due to subjects (some people give lower ratings than others) and forecast dates (some meteorological data sets are harder to describe than others), and then performing a one-way ANOVA on the normalized scores using generator as the independent variable.

The GLM analysis showed a very significant effect of generator on ratings, $p < 0.001$ (two-tailed). Table 3 shows the homogeneous subsets identified by the Tukey HSD post-hoc test. These correspond to the following pairwise results. For the experts, SUMTIME-Hybrid was significantly better than *p*CRU-random and *p*CRU-2gram, and *p*CRU-greedy was better than *p*CRU-random. For the non-experts, all systems were better than *p*CRU-random, and SUMTIME-Hybrid was also better than *p*CRU-2gram.

The SPSS GLM analysis also showed that scores were significantly affected by subject, for both experts and non-experts ($p < 0.001$); in other words, different individuals rated texts differently. Non-expert scores were also significantly influenced by forecast

Table 3
Experiment 1: Mean human ratings (single criterion of output quality), and homogeneous subsets from Tukey HSD analysis.

	Experts					Non-Experts			
	Mean	Subsets				Mean	Subsets		
SUMTIME-Hybrid	3.82	A			SUMTIME-Hybrid	3.90	A		
*p*CRU-greedy	3.59	A	B		*SUMTIME-Corpus*	3.62	A	B	
SUMTIME-Corpus	3.22	A	B	C	*p*CRU-greedy	3.51	A	B	
*p*CRU-roulette	3.11	A	B	C	*p*CRU-roulette	3.49	A	B	
*p*CRU-2gram	2.68		B	C	*p*CRU-2gram	3.29		B	
*p*CRU-random	2.43			C	*p*CRU-random	2.51			C

The subsets show which differences are statistically significant. More specifically, systems which are in the same subset *do not* have statistically significant differences in their mean ratings; systems which are not in the same subset *do* have statistically significant differences in their mean ratings. For example, for both experts and non-experts, *p*CRU-greedy is not significantly different from SUMTIME-Hybrid (since both are in subset A), and *p*CRU-greedy is not significantly different from *p*CRU-2gram (since both are in subset B). However, SUMTIME-Hybrid is significantly different from *p*CRU-2gram, since no subset contains both of these.

date ($p < 0.001$); in other words, some forecast data sets were harder to describe than others.

The fact that subject and forecast date influence human ratings suggests that a Latin Square experimental design should be used. In a Latin Square design, every subject rates the same number of texts from each generator, and every generator is rated an equal number of times on each forecast date; this reduces the impact of idiosyncratic differences between individual subjects (and forecast dates). For example, in Experiment 1, expert subject TR gave higher ratings than average (mean of 4.17, against an overall mean for expert subjects of 3.06), as did non-expert subject MP (mean of 4.57, again an overall mean for non-experts of 3.34). In the non-Latin-Square design used for expert subjects, TR rated 4 texts from *p*CRU-greedy, but only one text from SUMTIME-Hybrid; this may have inflated *p*CRU-greedy's mean score (Table 3) relative to SUMTIME-Hybrid's. In the Latin Square design used for non-experts, in contrast, all the subjects, including MP, rated the same number of texts (three) from each generator; hence MP's generosity presumably benefited all generators equally, and did not inflate the score of any one generator compared to the other generators.

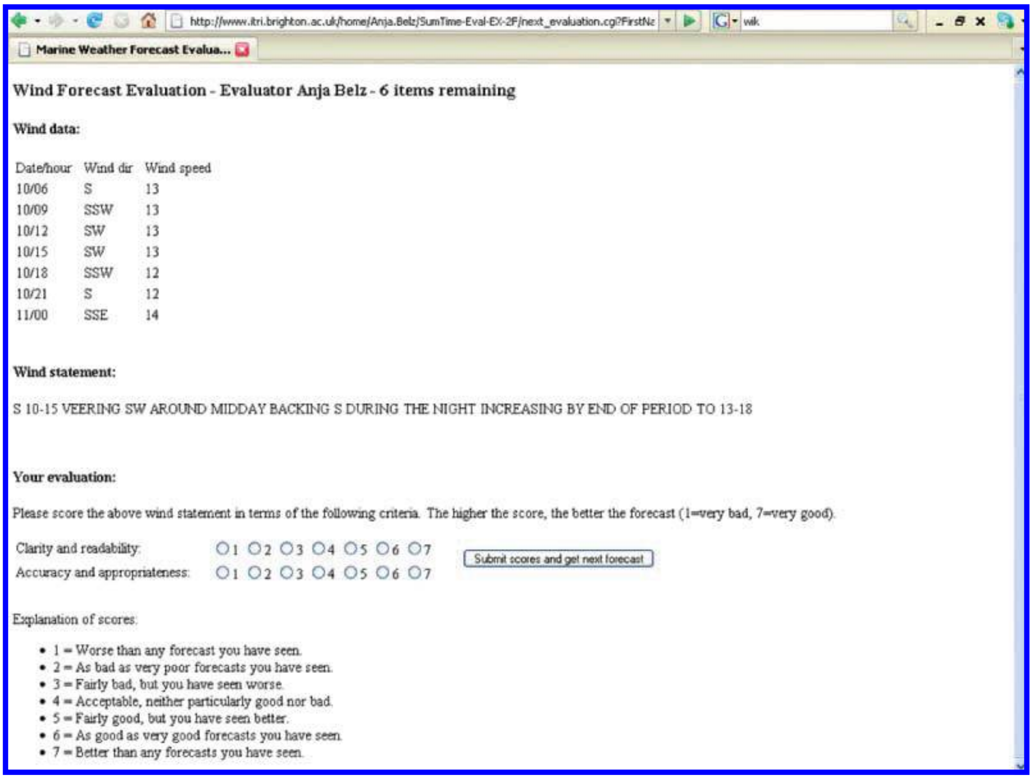
3.2.2 Second Human Evaluation.

Our first experiment focused on linguistic expression, but of course content determination is very important in NLG, so we decided to run another experiment which also included texts generated from meteorological data, by systems which performed content determination (that is, SUMTIME and Template). In this experiment we showed subjects the raw forecast data and asked them for separate ratings on "clarity and readability" (which was intended to elicit an assessment of linguistic quality) and "accuracy and appropriateness" (which was intended to elicit an assessment of content quality). For brevity, we refer to these scores as Clarity and Accuracy, respectively, herein.

We used 14 new randomly selected forecast dates, and 14 new expert subjects (we did not ask non-experts to rate these texts, because we were not confident that they could assess the accuracy and appropriateness of texts). The subjects were asked to rate seven types of texts: corpus texts, SUMTIME texts, Template texts, and texts produced by the Experiment 1 systems (except that we dropped *p*CRU-2gram); we did not in this experiment ask subjects to rate reference texts. We would have liked to recruit more than 14 subjects, but this proved difficult (see also Section 4.2); however, 14 subjects is an improvement over the 9 expert subjects used in Experiment 1 from the perspective of limiting the impact of individual differences between subjects.

We also made a number of changes to our experimental design, based on issues identified in the first experiment with expert subjects. The most important ones were that we used a Latin Square design (with two subjects rating each system/date combination), we asked for ratings on a seven-point scale instead of a six-point one (so the scale had a middle position which subjects could select), we explicitly gave instructions as to what the ratings meant (to reduce variation due to differing interpretations of the scale), and we carried out a non-parametric as well as parametric statistical analysis. A screenshot from the experiment is shown in Figure 5.

There was a significant ($p < 0.001$) correlation between the accuracy and clarity scores that subjects gave to texts (Pearson $r = 0.58$), when computed on the 196 individual ratings made by subjects (when correlation is computed on the mean values, significance cannot be shown, because there are far fewer data points: Pearson's $r = 0.572$, $p = 0.09$). It is not clear whether this is because subjects did not properly distinguish accuracy from clarity, or because generators that generated high-accuracy texts (such as SUMTIME) also generated high-clarity texts.


http://www.itri.brighton.ac.uk/home/Anja.Belz/SumTime-Eval-EX-2F/next_evaluation.cgi?FirstNa
wik
Marine Weather Forecast Evalua...

Wind Forecast Evaluation - Evaluator Anja Belz - 6 items remaining

Wind data:

Date/hour	Wind dir	Wind speed
10/06	S	13
10/09	SSW	13
10/12	SW	13
10/15	SW	13
10/18	SSW	12
10/21	S	12
11/00	SSE	14

Wind statement:

S 10-15 VEERING SW AROUND MIDDAY BACKING S DURING THE NIGHT INCREASING BY END OF PERIOD TO 13-18

Your evaluation:

Please score the above wind statement in terms of the following criteria. The higher the score, the better the forecast (1=very bad, 7=very good).

Clarity and readability: ○1 ○2 ○3 ○4 ○5 ○6 ○7

Accuracy and appropriateness: ○1 ○2 ○3 ○4 ○5 ○6 ○7

Submit scores and get next forecast

Explanation of scores:

- 1 = Worse than any forecast you have seen.
- 2 = As bad as very poor forecasts you have seen.
- 3 = Fairly bad, but you have seen worse.
- 4 = Acceptable, neither particularly good nor bad.
- 5 = Fairly good, but you have seen better.
- 6 = As good as very good forecasts you have seen.
- 7 = Better than any forecasts you have seen.

Figure 5
Screenshot from Experiment 2.

The *averaged* results of the human evaluations in Experiment 2 are shown in Table 4. Because the Experiment 1 texts were communicating the same content, and only differed in linguistic expression, it seems likely that Experiment 2's clarity scores should correlate with Experiment 1's scores. This is indeed the case: The correlation between average scores for the five systems that were included in both experiments is high with Pearson's $r = 0.9$ (p < 0.05).

Table 4
Experiment 2: Clarity and Accuracy average scores (expert subjects); homogeneous subsets from Tukey HSD analysis.

	Clarity					Accuracy		
	Mean	Subsets				Mean	Subsets	
SUMTIME	5.04	A			SUMTIME	5.07	A	
*p*CRU-greedy	4.79	A	B		Template	4.46	A	B
SUMTIME-Hybrid	4.61	A	B		*p*CRU-greedy	4.32	A	B
*p*CRU-roulette	4.54	A	B		*SUMTIME-Corpus*	4.11		B
SUMTIME-Corpus	4.50	A	B		SUMTIME-Hybrid	3.96		B
Template	4.04		B	C	*p*CRU-roulette	3.89		B
*p*CRU-random	3.64			C	*p*CRU-random	3.79		B

The SPSS GLM found a very significant effect of generator on both accuracy and clarity scores ($p < 0.001$). Table 4 shows the homogeneous subsets identified by the Tukey HSD post hoc test. The corresponding pairwise results are as follows. For the clarity scores, SUMTIME is significantly better than Template and *p*CRU-random; and all systems except Template are better than *p*CRU-random. For the accuracy scores, SUMTIME is significantly better than all systems except *p*CRU-greedy and Template. The GLM analysis also showed that subject and forecast date (as well as generator) had a significant impact on accuracy and clarity ratings ($p < 0.002$).

This statistical analysis assumes that it is appropriate to use ANOVA-like tests to analyze quality ratings. Although this is common practice in many NLP papers, including most previous validation studies of automatic metrics which we are aware of, a good argument can be made that quality ratings should be analyzed using non-parametric tests. This is because ratings are ordinal, so it is not clear that it makes sense to compute their mean, which is what the ANOVA and GLM tests do. Hence we also carried out a non-parametric analysis to identify significant differences in the human ratings. More specifically, we used the Wilcoxon Signed-Rank test, with a Bonferroni multiple-hypothesis correction, to identify pairs of systems that had significantly different ratings. When comparing two systems, the Wilcoxon test requires each rating of the first system to be paired with a related rating of the second system. Because we had two ratings from every subject for each system, we paired the lowest rating that a subject gave to a text produced by the first system with the lowest rating that that subject gave to a text produced by the second system; we similarly paired the highest ratings given by each subject to the two systems.

For example, subject AG evaluated two SUMTIME texts and gave them clarity ratings of 5 and 6; he also evaluated two Template texts, and gave them clarity ratings of 4 and 6. In our non-parametric analysis, we paired the lowest rating given by AG to a SUMTIME text (that is, 5) with the lowest rating given by AG to a Template text (that is, 4); we also paired the highest rating given by AG to a SUMTIME text (that is, 6) with the highest rating given by AG to a Template text (that is, 6). This pairing was possible because we used a Latin Square design in this experiment, which meant that every subject rated the same number of texts (two) from each generator.

This procedure identified the same four significant differences in Accuracy as in Table 4—namely, SUMTIME is significantly better than all systems except *p*CRU-greedy and Template. However, it identified only three significant differences in Clarity—namely, SUMTIME is significantly better than Template and *p*CRU-random, and *p*CRU-greedy is better than *p*CRU-random (this is a subset of the significant differences in Clarity shown in Table 4).

3.3 Correlation between Automatic Metrics and Human Judgments

In line with standard practice in validating metrics in MT, our main tool in analyzing the ability of automatically calculated metrics to predict human judgments is calculating Pearson correlation coefficients between sets of scores produced by metrics and the human ratings from Section 3.2. We computed correlations between metric scores for the different systems and the mean human ratings for these systems; for example, we computed the correlation between the BLEU score for SUMTIME and the average rating given by the human subjects to SUMTIME texts. We did not compute correlations on individual texts; for example, we did not try to correlate the BLEU score for the specific SUMTIME text shown in Table 2 against the human ratings of this specific text. This is because metrics such as BLEU and ROUGE are not intended to be meaningful for individual sentences.

Note that because correlations are being computed on a small set of numbers (seven at most), fairly high correlation coefficients are needed to achieve significance. In part because of this, we were less conservative in our statistical significance calculations than in the experiments reported in Section 3.2. In particular, we computed statistical significance of correlations using one-tailed (instead of two-tailed) tests, and we did not apply a Bonferroni multiple-hypothesis correction. None of the correlations presented subsequently would be significant if Bonferroni-adjusted two-tailed p-values were used; indeed we could only realistically expect to get significant correlations under this measure if we looked at more systems and/or fewer metrics (this is further discussed in Section 4.3).

3.3.1 Metrics and Reference Texts Used. We tested five automatic corpus-based metrics: two variants of the BLEU metric used in machine translation (Papineni et al. 2002); two variants of the ROUGE metric used in document summarization (Lin and Hovy 2003); and a simple sting-edit distance metric (as a baseline).

BLEU is a precision metric that assesses the quality of a generated text in terms of the proportion of its word *n*-grams that it shares with reference texts. BLEU scores range from 0 to 1, where 1 is the highest which can only be achieved by a generated text if all its substrings can be found in one of the reference texts (hence a reference text will always score 1). BLEU should be calculated on a large test set with multiple reference texts. We used BLEU-4[2] (that is, BLEU calculated using *n*-grams of size up to $n = 4$) because this version of BLEU is the main metric used in recent NIST Machine Translation evaluations (and indeed seems to have become a standard in the MT community). We also used the NIST[3] MT evaluation score (Doddington 2002); this is an adaptation of BLEU which gives more weight to less frequent *n*-grams which are assumed to be more informative.

There are several different ROUGE metrics. The simplest is ROUGE-*N*, which computes the highest proportion in any reference text of *n*-grams of length *N* that are matched by the generated text. A procedure is applied that averages the score across leave-one-out subsets of the set of reference texts. ROUGE-*N* is an almost straightforward *n*-gram recall metric between two texts, and has several counter-intuitive properties, including that even a text composed entirely of sentences from reference texts cannot score 1 (unless there is only one reference text). ROUGE-SU*N* looks at "skip bigrams" that occur in the generated text and reference texts; a skip bigram is two words which are not necessarily adjacent, but may be separated by up to *N* intermediate words. We used ROUGE-2 and ROUGE-SU4[4] because these are the main automatic metrics used in recent NIST Document Understanding Conferences (DUC).

We also included string-edit distance as a very simple automatic metric, which can be considered a sort of baseline. String-edit distance (SE) was computed with substitution at cost 2, and deletion and insertion at cost 1, and normalized to range 0 to 1 (perfect match). When multiple reference texts are used, the SE score for a generated text is the average of its scores against the reference texts; the SE score for a set of generated texts is the average of scores for the individual texts.

[2] Calculated by `ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-v11b.pl`.

[3] Calculated by `http://cio.nist.gov/esd/emaildir/lists/mt_list/bin00000.bin`.

[4] ROUGE code obtained via `http://www.isi.edu/~cyl/ROUGE/latest.html`.

Table 5
Experiment 1: Metric scores against three reference texts (produced by rewriting corpus texts), for the set of 18 forecasts used in expert evaluation.

System	Exp	Non	NIST-5	BLEU-4	ROUGE-SU4	ROUGE-2	SE
ST-Hybrid	3.82	3.90 (1)	6.382 (3)	0.584 (4)	0.558 (4)	0.528 (4)	0.705 (5)
*p*CRU-greedy	3.59	3.51 (3)	6.871 (2)	0.694 (2)	0.656 (2)	0.634 (2)	0.800 (2)
ST-Corpus	*3.22*	*3.62 (2)*	*8.705 (1)*	*0.951 (1)*	*0.839 (1)*	*0.815 (1)*	*0.917 (1)*
*p*CRU-roulette	3.11	3.49 (4)	6.206 (4)	0.563 (5)	0.554 (5)	0.51 (5)	0.735 (4)
*p*CRU-2gram	2.68	3.29 (5)	5.925 (5)	0.598 (3)	0.586 (3)	0.556 (3)	0.783 (3)
*p*CRU-random	2.43	2.51 (6)	4.608 (6)	0.355 (6)	0.462 (6)	0.419 (6)	0.649 (6)

For convenience, expert (Exp) and non-expert (Non) scores are also shown.

All of these automatic metrics require reference texts. For Experiment 1, which focused on content-to-text, we asked three meteorologists[5] (who had not contributed to the original SUMTIME-METEO corpus) to rewrite the corpus texts for the 21 dates used in Experiment 1 (each meteorologist rewrote all 21 corpus texts), correcting and improving them as they saw fit; examples are shown in Table 1. In principle, it would have been preferable to ask the forecasters to write texts based on content tuples (the actual input to the systems), but this is not a natural task for forecasters (they write texts from data, not from content tuples). However, asking them to rewrite corpus texts meant they were unable to change the content, and focused on lexical choice and syntactic structure, as intended.

For Experiment 2, which also looked at systems which performed content determination, we asked the same three meteorologists to write reference texts based on the raw numerical input data for the 14 dates used in Experiment 2 (each meteorologist wrote a text for all 14 dates); examples are shown in Table 2. They were not shown the corpus forecasts corresponding to the data.

3.3.2 Correlation with Results from Experiment 1. Table 5 shows the average scores each metric assigned to each system when calculated on the texts used in the Experiment 1 expert evaluations.[6] The metrics all rank the corpus texts highest, the *p*CRU-greedy texts second, and the *p*CRU-random texts lowest. Their strong preference for the corpus texts is probably an artefact of the way the reference texts were produced. The forecasters were asked to rewrite the corpus texts, which resulted in considerable similarity between the reference texts and the corpus texts. In calculating correlation figures (shown in Table 6), we therefore produced two sets of figures, one for the NLG systems and the corpus texts (I in the table) and one for just the NLG systems (II); set II should be regarded as a post hoc analysis. For set I, none of the metrics significantly correlate

[5] When we first reported results for Experiment 1 in Belz and Reiter (2006), we only had reference texts from two meteorologists, but we have since obtained reference texts from a third meteorologist. This is why the numbers in Tables 5 and 6 differ from the numbers given in Belz and Reiter (2006).

[6] The important information in Table 5 is the differences in the scores assigned by the same metric to different systems. Differences in the scores assigned by different metrics to the same system are not meaningful; they are just mathematical artefacts of the formulas used to calculate the metrics. For example, the fact that BLEU-4 gives SUMTIME-Hybrid a higher score than *p*CRU-random is important: this shows that BLEU-4 (correctly) predicts that human subjects prefer SUMTIME-Hybrid texts to *p*CRU-random texts. The fact that String-Edit (SE) gives a higher rating than BLEU-4 to SUMTIME-Hybrid is not important, as it does not tell us anything about how well SE or BLEU-4 can predict differences in human ratings of texts.

Table 6
Experiment 1: Correlation (Pearson's *r*) between human scores and automatic metrics.

	Experts	Non-exp.	NIST-5	BLEU-4	ROUGE-SU4	ROUGE-2	SE
I. Pearson's r, all NLG systems and corpus texts							
Experts	1	**0.874**	0.534	0.461	0.362	0.379	0.260
Non-exp.	**0.874**	1	0.685	0.639	0.518	0.527	0.483
II. Pearson's r, just NLG systems (not corpus texts) (post hoc)							
Experts	1	**0.884**	**0.836**	0.700	0.611	0.616	0.339
Non-exp.	**0.884**	1	**0.886**	0.797	0.654	0.647	0.505

Significant correlations (1-tailed, no Bonferroni correction) are shown in **bold**.

with human ratings. For set II, NIST-5 has a significant correlation with both expert and non-expert scores.

3.3.3 Correlation with Results from Experiment 2. Table 7 shows the average scores each metric assigned to each system for the texts used in Experiment 2: The rankings assigned by NIST-5 and BLEU-4 are identical, and SE largely agrees, with small differences in the top rankings (where differences between scores are very small), whereas the ROUGE metrics disagree with the other metrics to some degree. Table 8 shows the corresponding correlation figures for all NLG systems and the corpus texts (I), all NLG systems (II), and texts that communicate the same content (III). Set III consists of corpus texts and texts produced by systems whose input is corpus-derived content tuples (*p*CRU and SumTime-Hybrid). Note that the figures in sets I and II are similar; because reference texts for Experiment 2 were produced from raw data, the automatic metrics are not biased towards the corpus texts as they were in Experiment 1.

The most striking result from the correlation figures is that not a single metric correlates significantly with human judgments of accuracy. Clarity is significantly correlated with NIST-5 in sets I and III, and BLEU-4 and SE in set III.

This analysis assumes that it is sensible to use mean values of the human ratings; however, as mentioned at the end of Section 3.2.2, a good argument can be made that ordinal ratings should not be averaged. An alternative way of assessing the predictive accuracy of the metrics is to count how many of the significant differences identified by the non-parametric analysis in Section 3.2.2 were predicted by differences in metric scores. For example, the non-parametric analysis showed that human subjects

Table 7
Experiment 2: Metric scores against three reference texts (written from raw data), for the set of 14 forecasts.

System	Cla.	Acc.	NIST-5	BLEU-4	ROUGE-SU4	ROUGE-2	SE
SumTime	5.04	5.07 (1)	4.668 (5)	0.187 (5)	0.241 (5)	0.155 (6)	0.392 (5)
*p*CRU-greedy	4.79	4.32 (3)	5.118 (2)	0.244 (2)	0.258 (3)	0.180 (3)	0.42 (2)
ST-Hybrid	4.61	3.96 (5)	5.223 (1)	0.281 (1)	0.281 (1)	0.227 (1)	0.415 (3)
*p*CRU-roulette	4.54	3.89 (6)	4.798 (4)	0.221 (4)	0.244 (4)	0.169 (4)	0.421 (1)
ST-Corpus	4.50	4.11 (4)	4.969 (3)	0.227 (3)	0.281 (1)	0.21 (2)	0.415 (3)
Template	4.04	4.46 (2)	3.299 (7)	0.106 (7)	0.168 (7)	0.089 (7)	0.333 (7)
*p*CRU-random	3.64	3.79 (7)	4.011 (6)	0.162 (6)	0.238 (6)	0.156 (5)	0.379 (6)

For convenience, human ratings also shown.

Table 8
Experiment 2: Correlation (Pearson's *r*) between human score and metrics.

	Cla.	Acc.	NIST-5	BLEU-4	ROUGE-SU4	ROUGE-2	SE
I. Pearson's r, all NLG systems and corpus texts							
Clarity	1	0.572	**0.701**	0.577	0.431	0.401	0.571
Accuracy	0.572	1	−0.118	−0.281	−0.308	−0.375	−0.286
II. Pearson's r, just NLG systems (not corpus texts)							
Clarity	1	0.583	0.711	0.578	0.455	0.417	0.578
Accuracy	0.583	1	−0.092	−0.265	−0.285	−0.358	−0.266
III. Pearson's r, content-to-text NLG systems and corpus texts (same content)							
Clarity	1	0.741	**0.959**	**0.858**	0.550	0.549	**0.969**
Accuracy	0.741	1	0.682	0.502	0.432	0.294	0.607

Significant (1-tailed, no Bonferroni correction) correlations are shown in **bold**.

considered SumTime to be significantly more accurate than *p*CRU-random. If we look at the metric scores shown in Table 7, we see that ROUGE-2 rated *p*CRU-random higher than SumTime, and all other metrics rated SumTime higher than *p*CRU-random. Hence ROUGE-2 did not predict this significant difference in human Accuracy ratings (SumTime better than *p*CRU-random), but the other metrics did.

Table 9 shows the predictive accuracy of the metrics (in this sense). Overall this agrees with the main finding of the parametric analysis (see Table 3), namely that existing metrics are better at predicting Clarity than Accuracy.

It is interesting to compare our findings with Stent, Marge, and Singhai (2005), who examined the correlation between human judgments and several automatic metrics (including BLEU, NIST, and ROUGE) when evaluating computer-generated paraphrases. They in fact found more or less the opposite result, namely, that the metrics correlated with adequacy (similar to our Accuracy) but not fluency (similar to our Clarity). This may partially be due to the fact that Stent, Marge, and Singhai used a single reference text, which was the original input text to the paraphraser. With such a reference text, we wonder if the metrics largely measured the amount of paraphrasing done; that is, texts with less paraphrasing (whose surface form was thus closer to the original texts) were rated higher by the metrics. If so, it would not be surprising if the metrics correlated with accuracy but not with fluency, because it is possible that subjects regarded sentences whose surface form was close to the original text as more accurate but not necessarily more fluent. In their discussion section, Stent, Marge, and Singhai acknowledged that using a single reference text is problematic, and recommended that multiple reference

Table 9
Number of significant non-parametric differences predicted by each metric.

Metric	Clarity	Accuracy
NIST-5	3	1
BLEU-4	3	1
ROUGE-SU4	3	1
ROUGE-2	2	0
SE	3	1
actual number of significant diff	3	4

texts should be used if possible; however they also pointed out that this is not a panacea, since even 3–4 reference texts are unlikely to capture all of the acceptable variations in a text.

3.3.4 Summary of Results. In Experiment 1 all systems communicated the same content (they take the same content tuples as inputs), subjects were asked to give a single overall rating for texts, and reference texts were created by rewriting corpus texts. Under these conditions, NIST-5 scores are significantly correlated with expert scores (Table 6).

In Experiment 2, systems communicated different contents, subjects were asked to give separate clarity and accuracy ratings for a system, and reference texts were created by writing new texts from numerical data. Under these conditions, NIST-5 is significantly correlated with human clarity judgments. If only texts which communicate the same content at the content tuple level are included in the analysis, then NIST-5 and SE are strongly correlated with clarity judgments ($r > 0.95$), and BLEU-4 also correlates significantly (Table 8). However, no metric correlates significantly with human accuracy judgments under any analysis.

4. Discussion: What Can We Conclude from Our Results

The most obvious interpretation of our results is that it is acceptable to use BLEU-like metrics (with caution) to estimate the linguistic quality of generated texts, especially when comparing texts which are communicating the same content; but current automatic metrics should not be used to evaluate the quality of the content of generated texts. However, there are a number of caveats which must be considered which we discuss subsequently. These caveats may have relevance to other validation studies of automatic metrics in NLP.

Determining the validity of "cheap" evaluation techniques which are intended to approximate the genuine outcome measure is of course a problem that occurs in many areas of science, and we believe it is useful to look at what other fields do in this regard. Hence we relate our discussion to validation requirements in clinical medicine for "surrogate measures" (Greenhalgh 2006) (for example, using blood tests that measure HIV viral load to evaluate the effectiveness of AIDS treatments, instead of measuring actual mortality); and criterion validation requirements in psychology for psychometric and other tests (Kaplan and Saccuzzo 2001).

One general lesson from psychology is that there can be a strong temptation to use evaluation techniques which are quick, cheap, and appear to be impartial, even if they are known to have very limited validity. For example, psychometric tests which claim to predict academic success, such as the American SAT test, are very heavily used by American universities when they make admissions decisions, despite the fact that numerous validation studies have shown that these tests are poor predictors of how well a student does at university over the four-year span of a typical degree (although they do have a limited correlation with academic performance in a student's first year at university).

4.1 Generality across Domains, Genres, and Systems

Our experiments have been carried out in the specific domain and genre of marine weather forecasts for offshore oil rigs, and are based on a set of seven specific NLG systems. Will they apply to other domains, and indeed even to marine-weather-forecast generators built with different NLG technologies? Of course similar concerns have been raised about automatic metrics in other areas of NLP. For example, most validation studies of automatic metrics in machine translation and document summarization have been done with newswire texts; it is not clear, however, that results obtained from translated newswire texts also apply to translated scientific papers, for example.

A similar point is made strongly in the medical and psychological literature: Validation studies are performed in a particular context, and it is very risky to generalize them to other contexts, without additional evidence that they are effective in these new contexts. For example, Kaplan and Saccuzzo (2001, chapters 11 and 19) discuss the WAIS intelligence test, the original version of which was developed solely using data from subjects of European descent. Later research suggested it was not valid for other subjects; for example a variant called WISC, which was used in some school systems to decide which children should go to special education classes, was shown in the 1970s to correlate with teacher assessments of children of European descent, but not with teacher assessments of children from other ethnic groups. The test was subsequently revised to enhance its validity for children from diverse backgrounds.

We do not know how generalizable our findings are to other NLG contexts. We would have a better idea of generalizability if we performed similar experiments in other domains and genres, using systems built with a wide range of NLG technologies; but of course we cannot realistically expect to conduct enough experiments to examine *all* domains, genres, and technologies of interest.

Ultimately, the key to generalizing experimental results is a good theoretical model which is scientifically plausible and fits the experimental data. Perhaps for this reason, surrogate measures used in medicine are expected to be biologically plausible predictors of the actual outcome measures as well as empirically correlated with them (Greenhalgh 2006, page 95). The theoretical basis behind most current metrics used in NLG seems to be an intuition that similarity in surface forms should correlate with similarity in task effectiveness. It may be worth investigating whether psycholinguistic models of language comprehension (for example, Kintsch 1998) could provide a stronger theoretical basis for metric plausibility.

For what it is worth, our intuition is that our findings will apply to other application domains which involve generating texts which are short, linguistically simple, and not very varied. We would be extremely cautious about attempting to generalize our results to application domains which require texts which are longer, more complex, and more varied, such as BabyTalk.

4.2 Does Correlation with Human Judgments Mean Correlation with Task-Effectiveness?

As mentioned in Section 2.1.1, task-effectiveness evaluations are the most highly regarded evaluations in NLG; ultimately what we usually want to know is how effective NLG texts are in achieving their communicative goal, not whether readers like them or not. From this perspective, a major weakness in our study is that we correlated automatic ratings with human ratings, not task-effectiveness evaluations.

Attempts to correlate automatic metrics with task-based evaluations have been quite rare. The only ones we are aware of in NLG took place in the Generation Challenges events mentioned earlier; none of the automatic metrics used in these events had a significant correlation with task performance. In the summarization community, Dorr et al. (2005) found very weak correlation between an automatic metric (ROUGE) and task performance. We are not aware of any studies in machine translation which have analyzed correlation between automatic metrics and task-performance.

This is a major concern (as noted by Belz 2009) because we also do not know how well human ratings predict task-effectiveness; in other words, the fact that NIST scores predict human clarity ratings of NLG texts does not guarantee that NIST scores will predict task effectiveness, because we do not know that human clarity ratings correlate with task effectiveness. Looking again at medicine and psychology, validation studies in these fields need to show correlation with the actual outcome variable or at least a previously validated measure; in the words of Kaplan and Saccuzzo (2001, page 141), "a meaningless [test] which is well correlated with another meaningless [test] remains meaningless."

A major reason why so few correlation studies have been done between automatic metrics and task effectiveness is the significant amount of resources needed for such studies (and this is why we did not look at task-effectiveness in this study). The problem is not just time and money, it is also that task-based evaluations require support from domain experts (as mentioned in Section 2.1.1), and such support can be difficult to get for validation studies. To take a concrete example, a senior consultant at a hospital might be willing to encourage his medical colleagues to participate in the evaluation of a high-quality NLG system, for the purpose of determining whether this system was a useful medical decision-support aid; but such a consultant might be less willing to encourage his colleagues to participate in the evaluation of several NLG systems of mixed quality, for the purpose of determining whether human ratings correlated with automatic metrics.

Even obtaining subjects for ratings-based correlation studies can be difficult. For example, when we ran a human judgment-based study to test the effectiveness of SUMTIME texts (Reiter et al. 2005), we managed to recruit 72 subjects in a few weeks; in contrast it took us several months to recruit the 23 expert subjects who participated in the studies reported in this article. Both experiments required similar time commitments from similar subjects. However, subjects (and the domain experts who facilitated subject recruitment) were much more enthusiastic about testing the effectiveness of a system which they might themselves use; they were less enthused about testing hypotheses about correlations between NLG evaluation metrics.

A related issue is how well human judgments of the *clarity* of texts correlate with human judgments of the *overall quality* of texts; this is important because our results suggest that current metrics are much better at predicting human clarity judgments than human accuracy judgments. Intuitively, it seems likely that readers place more importance on content than on linguistic expression. In the SUMTIME domain, this intuition is supported by the SUMTIME evaluation (Reiter et al. 2005), in which forecast readers were asked to compare two forecasts, and say which was easier to read, which was more accurate, and which was overall more appropriate. In cases where subjects rated one forecast as easier to read and another as more accurate, they said the "more accurate" forecast was overall more appropriate in 55% of cases, and the "easier to read" forecast was overall more appropriate in only 18% of cases (in 27% of the cases they said neither of the forecasts was overall more appropriate than the other); this is significant at $p < 0.001$. Given this, it is a pity that the metrics we examined were so much better at predicting clarity than they were at predicting accuracy.

4.3 Statistical Issues

Like other scientific experiments, NLP evaluations are regarded as producing a significant result if they have a p-value (likelihood of incorrectly rejecting the null hypothesis) of 0.05 or less. Of course, statistical significance can be calculated in many ways; for example parametric or non-parametric tests can be applied, multiple-hypothesis (e.g., Bonferroni) corrections may or may not be applied, one or two-tailed p-values can be used, post hoc findings may or not be presented, and so on.

In medicine, recent work by Ioannidis (2005a, 2005b) and others (partially based on analyses of whether experimental results are replicated in follow-up studies) has suggested that a very conservative statistical analysis should be used in medical research. Ionnadis concludes that a very high quality medical experiment with very conservative statistical analysis has about an 85% chance of being replicated successfully; and that this chance quickly declines to noise levels once the design, execution, and/or statistical analysis of the experiment becomes less than ideal.

One could argue that computational linguistics should insist on similarly strict statistical analyses; in particular always use two-tailed p-values, always apply multiple hypothesis corrections, always discard post hoc findings (unless they are from tests specifically designed for post hoc analysis, such as Tukey HSD), and always use non-parametric tests if there is any doubt about the appropriateness of parametric tests. As we mentioned in Section 3.3, none of the correlations we observed between automatic metrics and human judgments would be considered significant under such a conservative statistical analysis. Indeed, in order to have a reasonable chance of seeing a statistically significant correlation under a conservative statistical analysis (to have sufficient *power* in a statistical sense), we would need to either look at more systems (since the value of Pearson's r needed to achieve statistically significant correlations decreases as the number of points in the correlation increase), and/or look at fewer metrics (since the impact of multiple hypothesis corrections decreases when fewer hypotheses are being tested).

The last point is particularly worth bearing in mind, because there is a strong temptation in validation studies to include as many metrics as possible. After all, once the human evaluations have been collected and the reference corpora have been created, we can compute correlations with other metrics (additional metrics such as METEOR (Banerjee and Lavie 2005), and variations of metrics we are already examining, such as BLEU-2 and BLEU-3 as well as BLEU-4) at the touch of a button. But if we are applying multiple hypothesis corrections, then there is a major drawback to including a large number of metrics in the study, which is that this will make it more difficult to find statistically significant correlations.

However, perhaps it is wrong to use such strict statistics in computational linguistics, and indeed we are aware of many reports in computational linguistics which present one-tailed p-values, do not apply multiple hypothesis corrections, present post hoc analyses as significant, and/or use parametric tests to analyse data which does not have the characteristics assumed by the parametric test. In this respect the practice in computational linguistics is perhaps closer to psychology, where (for example) multiple hypothesis corrections are less common than they are in medicine; indeed a textbook on statistics for psychologists used at the University of Aberdeen does not even mention the topic.

So should our results be considered statistically insignificant (using a very strict statistical analysis) or statistically significant (using a less strict statistical analysis)? Our personal opinion is that the correlations we have observed are real, but it would be extremely useful to verify this by running larger experiments which showed significant results under a stricter statistical analysis. However, other readers may have different opinions, and in this article we have tried to follow the advice of Greenhalgh (2006) by giving enough information about our statistical analyses to enable readers to make their own informed judgments as to how to interpret them.

5. Discussion: When Should Automatic Metrics be Used in Evaluating NLG?

Our goal in this experiment was to shed light on when automatic metrics should be used in NLG. Given all the previously mentioned caveats, we cannot of course draw firm conclusions about this topic. But we can make some suggestions.

First of all, the automatic metrics we examined should not be used to predict human judgments of content quality; none of them had a significant correlation with human accuracy judgments, even when statistical significance is calculated in a less-than-conservative fashion.

Second of all, even when evaluating linguistic quality, current automatic metrics should be used with caution, as a supplement rather than a replacement for human evaluation; similar comments have been made about the use of automatic metrics in MT (Papineni et al. 2002; Callison-Burch, Osborne, and Koehn 2006). Particular caution should be used when evaluating NLG systems which generate significantly longer and more complex texts than the marine weather forecasts we examined here. We are not aware of any validation studies on such texts, and there are important aspects of the linguistic quality of longer and more complex texts (such as discourse coherence) which are not measured by current metrics.

Thirdly, automatic metrics are most appealing in contexts where a suitable corpus of reference texts already exists. Creating good-quality reference texts is an expensive endeavor, especially in domains (such as summaries of clinical data) where texts must be written by skilled domain experts. Therefore we suspect that it may be difficult in many cases to justify creating large corpora of reference texts solely for the purposes of automatic evaluation of NLG systems.

5.1 Should Automatic Metrics be Used in Shared-Task Evaluations?

In the wider NLP community, automatic metrics are especially popular in shared-task evaluations. This is partially because such metrics have a very low marginal cost compared to human evaluations. Automatic metrics need reference texts, and obtaining good reference texts can be costly; but once a collection of reference texts has been created, it can be used to evaluate any number of systems. Also automatic metrics are very easy to use once the software and reference corpus has been created; developers do not need to be trained in using BLEU and ROUGE. In contrast human-based evaluations generally need a certain number of subjects *per system*, so their cost goes up with the number of systems evaluated. Also, expertise and/or training is needed to conduct experiments with people, which not all NLP researchers possess. Finally, evaluation with metrics is entirely reproducible.

However, despite the cost-effectiveness and other appealing aspects of automatic metrics in shared tasks, we do not believe that shared tasks in NLG should use automatic metrics as the sole evaluation criterion. Until there is better evidence that automatic metrics correlate with human evaluations, shared tasks in NLG should also include human evaluations, preferably task-effectiveness ones. This strategy is being followed in the Generation Challenges shared-task NLG events (Section 2.1.4).

5.2 Should Automatic Metrics be Used in Diagnostic Evaluation?

We have focused in this article on evaluations that measure the quality of generated texts, but many NLG developers are also interested in *diagnostic evaluations* whose purpose is to identify problems in a system and suggest improvements. From this perspective, an advantage of human evaluations is that human subjects can be asked to make free-text comments on the texts that they see, and these comments are often extremely useful from a diagnostic perspective. On the other hand, an advantage of automatic metrics is that they allow developers to rapidly evaluate changes to systems and algorithms; indeed, some machine translation researchers use automatic metrics to automatically tune parameters without human intervention (Och 2003). However, as Och points out, this is only sensible if automatic metrics are known to be very accurate predictors of text quality. Because our results suggest that current automatic metrics are not highly accurate predictors of the quality of texts produced by NLG systems, we recommend developers be cautious in using metrics for diagnostic evaluation, and do not use metrics for automatic parameter tuning.

On the other hand, automatic metrics do have a potential advantage in small diagnostic evaluations, which is that they are not influenced by the individual preferences of a small number of human subjects. There are large differences in how different human subjects rate texts, as we pointed out at the end of Section 3.2.1. Such differences are not unusual: we have seen them in most human evaluations of NLG systems which we have carried out. These differences can be controlled for in a large experiment which uses many subjects. But if a diagnostic evaluation is conducted with a small number of subjects, who are chosen partially on the basis of being easy to recruit, there is a risk that the preferences expressed by these subjects will not be representative of users in general, and hence may mislead the developer as to how the system should be changed.

6. Conclusions

Automatic evaluation metrics have many desirable properties, such as being fast, cheap, and repeatable, and they have had a significant impact in many areas of NLP. We have compared the scores produced by several popular metrics, including BLEU and ROUGE, to human evaluations of NLG systems, in the domain of weather forecasts. Our results suggest that it may be appropriate to use existing automatic metrics (with caution) to evaluate the linguistic quality of generated texts, especially if metric evaluations supplement (rather than replace) human evaluations; for example metric-based evaluations could be used to provide diagnostic feedback to developers in the period before a large human evaluation. NIST-5 is perhaps the best metric to use for this purpose (out of the ones we investigated). However, existing metrics should not be used to evaluate the content of texts. Also it would be premature to use metrics to test hypotheses about the effectiveness of NLG systems; we need more experimental validation data (including validation against task effectiveness measures) and ideally a good theoretical model as well.

Acknowledgments

Anja Belz's work was in part supported under UK EPSRC grant GR/S24480/01. Many thanks to Yaji Sripada for his help obtaining SUMTIME texts, and to Roger Evans, Chris Mellish, Kees van Deemter, and the anonymous reviewers for their very helpful comments. Last but not least, we would like to thank the 3 meteorologists, 23 expert forecast readers (master mariners for the most part), and 21 non-experts (family, friends, and colleagues) for their generous help in the evaluation experiments.

References

Banerjee, Satanjeev and Alon Lavie. 2005. Meteor: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*, pages 65–72, Ann Arbor, MI.

Bangalore, Srinivas, Owen Rambow, and Steve Whittaker. 2000. Evaluation metrics for generation. In *Proceedings of the 1st International Conference on Natural Language Generation*, pages 1–8, Mitzpe Ramon.

Belz, Anja. 2008. Automatic generation of weather forecast texts using comprehensive probabilistic generation-space models. *Natural Language Engineering*, 14:431–455.

Belz, Anja. 2009. That's nice ... what do you do with it? *Computational Linguistics*, 35:111–118.

Belz, Anja and Albert Gatt. 2007. The attribute selection for GRE challenge: Overview and evaluation results. In *Proceedings of the 2nd UCNLG Workshop: Language Generation and Machine Translation (UCNLG+MT)*, pages 75–83, Copenhagen.

Belz, Anja and Albert Gatt. 2008. Intrinsic vs. extrinsic evaluation measures for referring expression generation. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL'08)*, pages 197–200, Columbus, OH.

Belz, Anja and Ehud Reiter. 2006. Comparing automatic and human evaluation of NLG systems. In *Proceedings of the EACL'06*, pages 313–320, Trento.

Binsted, Kim, Helen Pain, and Graeme Ritchie. 1997. Children's evaluation of computer-generated punning riddles. *Pragmatics and Cognition*, 5:309–358.

Black, Rolf, Annalu Waller, Graeme Ritchie, Helen Pain, and Ruli Manurung. 2007. Evaluation of joke-creation software with children with complex communication needs. *Communication Matters*, 21:23–28.

Cahill, Aoife and Josef van Genabith. 2006. Robust PCFG-based generation using automatically acquired LFG approximations. In *Proceedings of ACL'06*, pages 1033–1044, Sydney.

Callaway, Charles. 2003. Evaluating coverage for large symbolic NLG grammars. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI 2003)*, pages 811–817, Acapulco.

Callison-Burch, Chris, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2008. Further meta-evaluation of machine translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 70–106, Columbus, OH.

Callison-Burch, Chris, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of BLEU in machine translation research. In *Proceedings of EACL-2006*, pages 249–256, Trento.

Carenini, Giuseppe and Johanna D. Moore. 2006. Generating and evaluating evaluative arguments. *Artificial Intelligence*, 170:925–952.

Coch, José. 1998. Interactive generation and knowledge administration in MultiMeteo. In *Proceedings of the Ninth International Workshop on Natural-Language Generation (INLG-1996)*, pages 300–303, Sussex.

Coughlin, Deborah. 2003. Correlating automated and human assessments of machine translation quality. In *Proceedings of MT Summit IX*, pages 63–70, New Orleans, LA.

Cunningham, Steven, Sarah Deere, Andrew Symon, Robert Elton, and Neil McIntosh. 1998. A randomized, controlled trial of computerized physiologic trend monitoring in an intensive care unit. *Critical Care Medicine*, 26:2053–2060.

Dang, Hoa. 2006. Duc 2005: Evaluation of question-focused summarization systems. In *Proceedings of the ACL-COLING 2006 Workshop on Task-Focused Summarization and Question Answering*, pages 48–55, Sydney.

Di Eugenio, Barbara, Michael Glass, and Michael Trolio. 2002. The DIAG experiments: Natural language generation for tutoring systems. In *Proceedings of the Second International Conference on Natural Language Generation (INLG-2002)*, pages 120–127, Harriman, NY.

Doddington, George. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the Second International Conference on Human Language Technology Research*, pages 138–145, San Diego, CA.

Dorr, Bonnie, Christof Monz, Stacy President, Richard Schwartz, and David Zajic. 2005. Methodology for extrinsic evaluation of text summarization: Does ROUGE correlate? *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 1–8, Ann Arbor, MI.

Flesch, Rudolf. 1949. *The Art of Readable Writing*. Harper Brothers, New York.

Gatt, Albert, Anja Belz, and Eric Kow. 2008. The TUNA Challenge 2008: Overview and evaluation results. In *Proceedings of the 5th International Natural Language Generation Conference (INLG'08)*, pages 198–206, Salt Fork, OH.

Gatt, Albert, Anja Belz, and Eric Kow. 2009. The TUNA-REG Challenge

2009: Overview and evaluation results. In *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG'09)*, pages 174–182, Athens.

Goldberg, Eli, Norbert Driedger, and Richard Kittredge. 1994. Using natural-language processing to produce weather forecasts. *IEEE Expert*, 9(2):45–53.

Greenhalgh, Trisha. 2006. *How to Read a Paper: The Basics of Evidence Based Medicine*. BMJ Books, Oxford, third edition.

Habash, Nizar. 2004. The use of a structural n-gram language model in generation-heavy hybrid machine translation. In *Proceedings of the 3rd International Conference on Natural Language Generation (INLG '04)*, volume 3123 of *LNAI*, pages 61–69, Brockenhurst.

Harris, Mary Dee. 2008. Building a large-scale commercial NLG system for am EMR. In *Proceedings of INLG-2008*, pages 157–160, Salt Fork, OH.

Hirschman, Lynette. 1998. The evolution of evaluation: Lessons from the message understanding conferences. *Computer Speech and Language*, 12:283–285.

Ioannidis, John. 2005a. Contradicted and initially stronger effects in highly cited clinical research. *Journal of American Medical Association*, 294:218–228.

Ioannidis, John. 2005b. Why most published research findings are false. *PLoS Medicine*, 2, doi:10.1371/*journal.pmed*.0020124.

Kaplan, Robert and Dennis Saccuzzo. 2001. *Psychological Testing: Principles, Applications, and Issues (Fifth Edition)*. Wadsworth, London.

Kintsch, Walter. 1998. *Comprehension*. Cambridge University Press.

Langkilde, Irene. 2002. An empirical verification of coverage and correctness for a general-purpose sentence generator. In *Proceedings of the 2nd International Natural Language Generation Conference (INLG '02)*, pages 17–24, Harriman, NY.

Langkilde, Irene and Kevin Knight. 1998. Generation that exploits corpus-based statistical knowledge. In *Proceedings of COLING-ACL 1998*, pages 704–710, Montreal.

Law, Anna, Yvonne Freer, Jim Hunter, Robert Logie, Neil McIntosh, and John Quinn. 2005. Generating textual summaries of graphical time series data to support medical decision making in the neonatal intensive care unit. *Journal of Clinical Monitoring and Computing*, 19:183–194.

Lester, James and Bruce Porter. 1997. Developing and empirically evaluating robust explanation generators: The KNIGHT experiments. *Computational Linguistics*, 23(1):65–101.

Lin, Chin-Ye and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of HLT-NAACL 2003*, pages 71–78, Edmonton.

Marciniak, Tomasz and Michael Strube. 2004. Classification-based generation using TAG. In *Natural Language Generation: Proceedings of INLG-2994*. Springer, pages 100–109, Brockenhurst.

Mellish, Chris and Robert Dale. 1998. Evaluation in the context of natural language generation. *Computer Speech and Language*, 12:349–373.

Nenkova, Ani and Rebecca Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. In *Proceedings of NAACL-2004*, pages 145–152, Boston, MA.

Oberlander, Jon. 1998. Do the right thing . . . but expect the unexpected. *Computational Linguistics*, 24:501–507.

Och, Franz Josef. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of ACL-2003*, pages 160–167, Sapporo.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of ACL-2002*, pages 311–318, Philadelphia, PA.

Portet, François, Ehud Reiter, Albert Gatt, Jim Hunter, Somayajulu Sripada, Yvonne Freer, and Cindy Sykes. 2009. Automatic generation of textual summaries from neonatal intensive care data. *Artificial Intelligence*, 173:789–816.

Reiter, Ehud, Roma Robertson, and Liesl Osman. 2003. Lessons from a failure: Generating tailored smoking cessation letters. *Artificial Intelligence*, 144:41–58.

Reiter, Ehud and Somayajulu Sripada. 2002. Should corpora texts be gold standards for NLG? In *Proceedings of the 2nd International Conference on Natural Language Generation*, pages 97–104, Harriman, NY.

Reiter, Ehud and Somayajulu Sripada. 2003. Learning the meaning and usage of time phrases from a parallel text-data corpus. In *Proceedings of the HLT-NAACL 2003 Workshop on Learning Word Meaning from Non-Linguistic Data*, pages 78–85, Edmonton.

Reiter, Ehud, Somayajulu Sripada, Jim Hunter, and Jin Yu. 2005. Choosing words in computer-generated weather forecasts. *Artificial Intelligence*, 167:137–169.

Reiter, Ehud, Somayajulu Sripada, and Roma Robertson. 2003. Acquiring correct knowledge for natural language generation. *Journal of Artificial Intelligence Research*, 18:491–516.

Scott, Donia and Johanna Moore. 2007. An NLG evaluation competition? Eight reasons to be cautious. In *Proceedings of the Workshop on Shared Tasks and Comparative Evaluation in Natural Language Generation*, pages 22–23, Arlington, VA.

Spärck Jones, K. and J. R. Galliers. 1995. *Evaluating Natural Language Processing Systems: An Analysis and Review*. Springer Verlag, Berlin.

Sripada, Somayajulu, Ehud Reiter, Ian Davy, and Kristian Nilssen. 2004. Lessons from deploying NLG technology for marine weather forecast text generation. In *Proceedings of PAIS-2004*, pages 760–764, Valencia.

Sripada, Somayajulu, Ehud Reiter, Jim Hunter, and Jin Yu. 2003. Exploiting a parallel TEXT-DATA corpus. In *Proceedings of Corpus Linguistics 2003*, pages 734–743, Lancaster.

Stent, Amanda, Matthew Marge, and Mohit Singhai. 2005. Evaluating evaluation methods for generation in the presence of variation. In *Proceedings of CICLing 2005*, pages 341–351, Mexico City.

van der Meulen, Marian, Robert Logie, Yvonne Freer, Cindy Sykes, Neil McIntosh, and Jim Hunter. 2009. When a graph is poorer than 100 words: A comparison of computerised natural language generation, human generated descriptions and graphical displays in neonatal intensive care. *Applied Cognitive Psychology*, doi:10.1002/acp.1545.

Walker, Marilyn, Owen Rambow, and Monica Rogati. 2002. Training a sentence planner for spoken dialogue using boosting. *Computer Speech and Language*, 16:409–433.

Williams, Sandra and Ehud Reiter. 2008. Generating basic skills reports for low-skilled readers. *Natural Language Engineering*, 14:495–535.

Young, Michael. 1999. Using Grice's maxim of quantity to select the content of plan descriptions. *Artificial Intelligence*, 115:215–256.

Zhong, Huayan and Amanda Stent. 2005. Building surface realizers automatically from corpora. In *Proceedings of UCNLG'05*, pages 49–54, Birmingham.